Published in: Schmidt, Sara/Siemens, Ray/Kumar, Amit/Unsworth, John (eds.): Digital Humanities 2007.
Conference Abstracts - Urbana-Champaign: University of Illinois, 2007. pp. 161-162.

Digital Text Resources for the Humanities – Legal Issues

Georg Rehm *(georg.rehm@uni-tuebingen.de)*
Tübingen University
Andreas Witt *(andreas.witt@uni-tuebingen.de)*
Tübingen University
Erhard Hinrichs *(eh@sfs.uni-tuebingen.de)*
Tübingen University
Timm Lehmberg
(timm.lehmberg@uni-hamburg.de)
Hamburg University
Christian Chiarcos
(chiarcos@ling.uni-potsdam.de)
Potsdam University
Felix Zimmermann *(mail@felix-zimmermann.eu)*
Institute for Legal Informatics
Hannover University
Heike Zinsmeister
(heike.zinsmeister@uni-tuebingen.de)
Tübingen University
Johannes Dellert *(jdellert@sfs.uni-tuebingen.de)*
Tübingen University

The session "Digital Text Resources for the Humanities – Legal Issues" consists of three papers that address the legal aspects connected to several crucial phases of handling text resources: collecting, compiling, curating, analysing, distributing, and archiving text resources such as corpora, are tasks carried out on a day-to-day basis by people involved in fields such as, for example, humanities computing, computational and corpus linguistics, information retrieval and text mining. Despite the ubiquity of document collections, the legal issues that are intrinsically tied to virtually all texts created and published by third parties (most importantly, their copyright, as well as privacy issues), do not typically attract a lot of interest. Though these issues are acknowledged, they are often regarded as rather insignificant for the research question at hand, or a project does not have any jurisprudential expertise to deal with legal issues in an adequate way. As a consequence, distributing a corpus (for example, to other interested

researchers) whose provenance is unknown or questionable, or publishing excerpts from a document collection on a website, may become next to impossible from a legal point of view. This is why scholars often decide not to publish their collections (or parts thereof) online at all, in order to avoid any potential legal problems. The session aims to provide an overview of the following legal aspects:

- The first contribution, "Language Corpora – Copyright – Data Protection: The Legal Point of View" (Timm Lehmberg, and Felix Zimmermann), highlights the legal requirements that hold with regard to the construction of digital text resources, special emphasis is given to the aspect of copyright and data protection (for example, potential reasons for the need to anonymise text corpora).
- The second presentation, "Collecting Legally Relevant Metadata by Means of a Decision-Tree-Based Questionnaire System" (Timm Lehmberg, Christian Chiarcos, Erhard Hinrichs, Georg Rehm, and Andreas Witt), consists of two parts: first, a web-based questionnaire is introduced that was developed to capture the requirements research projects have with regard to the archiving and distribution of their corpora; second, initial results from a study that spans three large research centres and more than 60 individual research projects are reported.
- The final paper, "Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections" (Georg Rehm, Andreas Witt, Heike Zinsmeister, and Johannes Dellert), introduces the idea of masking an annotated text corpus whose original source text collection is copyright-protected, so that the masked version can be distributed without any restrictions; furthermore, a fully working tool for masking an XML-annotated corpus is presented.

The authors of the three papers are associated with a joint project situated in three Collaborative Research Centres (SFB, Sonderforschungsbereich) that are sponsored by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft): SFB 441 (*Linguistic Data Structures*, Tübingen University), SFB 538 (*Multilingualism*, Hamburg University), and SFB 632 (*Information Structure*, Potsdam University). Each of these three research centres consists of about 15 to 20 research projects. Most projects work with digital text collections, in practically all cases these collections and corpora are constructed by the respective researchers themselves. A problem people involved in the fields of digital humanities or computational linguistics are often confronted with concerns the fact that the sustainability and reusability of corpora is not given too much attention – or that these aspects, in a worst case scenario, are completely ignored. Corpora are often created for an application or for a project that has a very specific research question, but when the project is finished it becomes next to impossible (especially for third parties) to gain access to the resource that took several months or maybe even years to create. The joint project *Sustainability of Linguistic Data* was therefore established to provide the conceptual, technical and infrastructural basis for a solution to the problem of sustainably archiving these digital text collections, addressing issues as diverse as, for example, annotation and metadata frameworks, best practice guidelines, legal issues of distributing text collections, and unifying diverse tag sets by means of an ontology.

Session Chairs: Georg Rehm and Andreas Witt